天玑舆情系统概述

翟立东 余智华 丁国栋

摘要: 随着互联网发展的日益深入，网络已成为当今社会信息传播的主要媒介之一。网络舆情形成迅速，对社会影响巨大，加强网络舆情的监测和分析，值得引起社会各界的高度重视。天玑舆情监测系统是专门针对网络舆情监测的工作要求和特点而打造的，不仅采用了专业化的搜索引擎技术，还融入了更加智能的数据挖掘技术，建立一个以日为周期的网络舆情监测平台，同时配上以周或者以月为基础的舆情分析报告，从而提供了一个便捷、科学、可操作的舆情工作平台。

关键词: 舆情；信息获取；分类聚类；数据挖掘

1 引言

所谓网络舆情指的是在一定的社会空间内，在网络中围绕社会事件的发生、发展和变化，民众对公共问题和社会管理者产生和持有的社会政治态度、信念和价值观。它是较多民众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和。

网络舆情呈现的影响力日趋增加，已渗入到从政府决策到社会政治、经济、文化和日常生活的各个层面，但种种偏颇的评论或损害国家安全与社会稳定的虚假信息也随之而来，正确引导网络舆情已成为维护社会安定的重大课题。

本文第二部分概述当前网络舆情的情况；第三部分分析了网络舆情监测的需求；第四部分深入浅出地对天玑舆情系统解决方案进行了介绍；最后总结天玑舆情系统的技术特点和实际效果。

2 网络舆情现状及特点

网络舆情所面向的是网络舆论这种新兴的舆论形式，它具有其它舆论形式所不具备的特点：

第一，网络舆情具有匿名性和虚拟性，难以有效监管和规范，发言者基本无需考虑承担法律责任的问题。

第二，网络信息真实性难以辨别，部分网民很少质疑信息的真实性，也没有进行成熟的思辨，仅凭主观臆断对信息发表意见，带有很浓厚的感情色彩，易导致真相的掩盖、言论的偏颇。

第三，网络舆情传播快、范围广、聚集效应强，但可控性较差。对在网络中四处粘贴攻击诋毁的消息，有关部门只能被动删除，防不胜防。

由于网络舆情的特点，使其能够产生巨大的引导效应，正面来看：

1. 网络舆情有利于真实民意的充分表达，使政府及时准确了解群众心声。由于网络民意生存空间的虚拟性，使得民意主体敢于表达自己的真实心声，在网络公共空间、尤其是网络论坛，网民能够就时事发表看法、展开争论。
2. 网络舆情能为政府科学决策提供依据，有利于扩大公众的参与权，提高公民的主人公地位。从整体上说，无论是哪一种网络舆情的出现，均体现了网民对国家的前途命运和社会的公共事务的关心，也体现了网民参与意识和主人翁精神的日益增强。在信息时代，网络舆情逐渐成为政府倾听民声、了解民意的一个重要渠道。
3. 通过意见领袖的引领能冲淡一些情绪型过激言论，有利于缓解渐趋紧张的社会矛盾。在网络论坛中有一大批的“舆论领袖”，他们的文字表达能力强、分析问题深刻、有独特见解，他们的言论往往在潜移默化中影响其他网民的看法。因此，由他们所引导的健康、理性的言论能控制一些负面的影响，有事半功倍的效果。
4. 网络舆情能促进对政府官员的监督、提高社会的透明度，有利于约束不良之风。网民参与的普遍性和不受控制性，使得网络舆情无时不在、无处不在，俨然一张群众监督的“天网”，使很多公共权力的运作被置于阳光之下，从而有利于促进社会透明度的增加和政府信息公开。

但是另一方面，网络舆情也会带来一些负面效应：

1. 网民的情绪化导致网络暴力频繁出现。由于发言者身份隐蔽，并且缺少规则限制和有效监督，有些人在面对困难和问题时，会把网络这个“虚拟”的世界作为不良情绪的宣泄空间。
2. 网络“把关人”缺失导致网络虚假信息泛滥，干扰了网民的正确判断，甚至于扰乱了正常的社会秩序。
3. “意见领袖”的恶意引导引发“蝴蝶效应”。一些反动、分裂势力在网络上以各种面目出现，制造和利用网络谣言，煽动激进情绪，使网络舆情显得异常复杂。

3 网络舆情监测需求

网络舆情的特点决定了舆情信息工作的时效性非常强。舆情变化的节拍经常是以小时计算。为了汇总舆情，要浏览和查找海量的网络信息，包括网络新闻报道、相关评论、网络论坛、博客等，从这些信息中提取与事件相关的舆情信息，然后分析舆情信息的时间与空间分布情况。为提高舆情工作的时效，必须充分运用现代网络技术，及时有效地进行信息搜集、信息处理、信息研判、信息反馈、决策，这就需要强有力的技术手段的支持。

从业务需求上看，网络舆情监测包括日常监测和突发事件监测两种方式：

日常监测，指将网络舆情监测作为本部门的一项日常工作不间断进行，随时掌握网络舆论的导向、特点和趋势。日常监测的意义在于，随时了解网络舆论的动态、方向，一旦发现负面的、重大的虚假舆情苗头，可以及时采取措施，对日常舆情进行引导，为有关部门提供决策支持。

突发事件监测，指当发生群体性突发事件时，对相关网络舆情的监测。突发事件的变化因素多，内部关系较为复杂，发展趋势难以预测，相关信息纷繁复杂，会给信息判断和决策增加很大的难度。另外，由于突发事件中的矛盾双方往往处于对立状态，影响或阻碍了原有信息沟通渠道的正常功能，从而给各种“小道消息”提供了填补信息真空的机会。此类事件突发性强、社会影响大、给决策者思考的时间短，如果不及时准确获得最新信息并加以判断处理，产生的后果非常严重。因此，在突发事件出现时，完善的舆情监测机制、及时有效的

舆情信息汇集和分析，全面掌握与该事件密切相关的各种信息，极其重要。

通过建立较为成熟的网络舆情日常监测和突发应对机制，尽量把负面舆情在苗头阶段就加以引导控制，当遇到重大突发事件时，就能够在短时间内调动和整合各种力量，形成联动，产生危机应对的合力。危机事件后能进行有效评估，包括危机情况、采取措施、对下一阶段走向的研判、对前一阶段应对的总结、反思与建议等，从而持续提升网络舆情监测和危机应对的能力。

网络舆情监测需要在互联网的海量信息中进行，工作时效性要求非常高，仅靠人工浏览的方法很难应对网上海量信息的收集和处理。这就需要现代信息网络技术、尤其是自动化的计算机软件系统的支持。

在相关的技术工具当中，搜索引擎提供了通用信息的检索服务，可以对舆情工作起到一定的帮助。其主要缺陷在于数据收录的及时性、深入性受到天然的限制，难以准确符合用户的需要。有价值的舆情线索往往被淹没在大量的搜索结果中，而且没有提供信息的统计分析，无法及时满足舆情监测对深入分析的要求。同时，搜索引擎的结果受主管部门、政策法规、商业利益所干预，需要的结果往往搜索不到，也会给舆情监测带来盲点。垂直搜索引擎，如新闻搜索、论坛搜索、生活搜索、财经搜索、图片搜索等，虽然信息更新比较及时，但仍然无法提供舆情信息的统计关联分析，更无法进行多人的协同工作与资源共享。日常监测尚且压力重重，遇到突发事件，更无法整合各种力量形成合力，难以应对网络危机。

4　天玑舆情系统解决方案

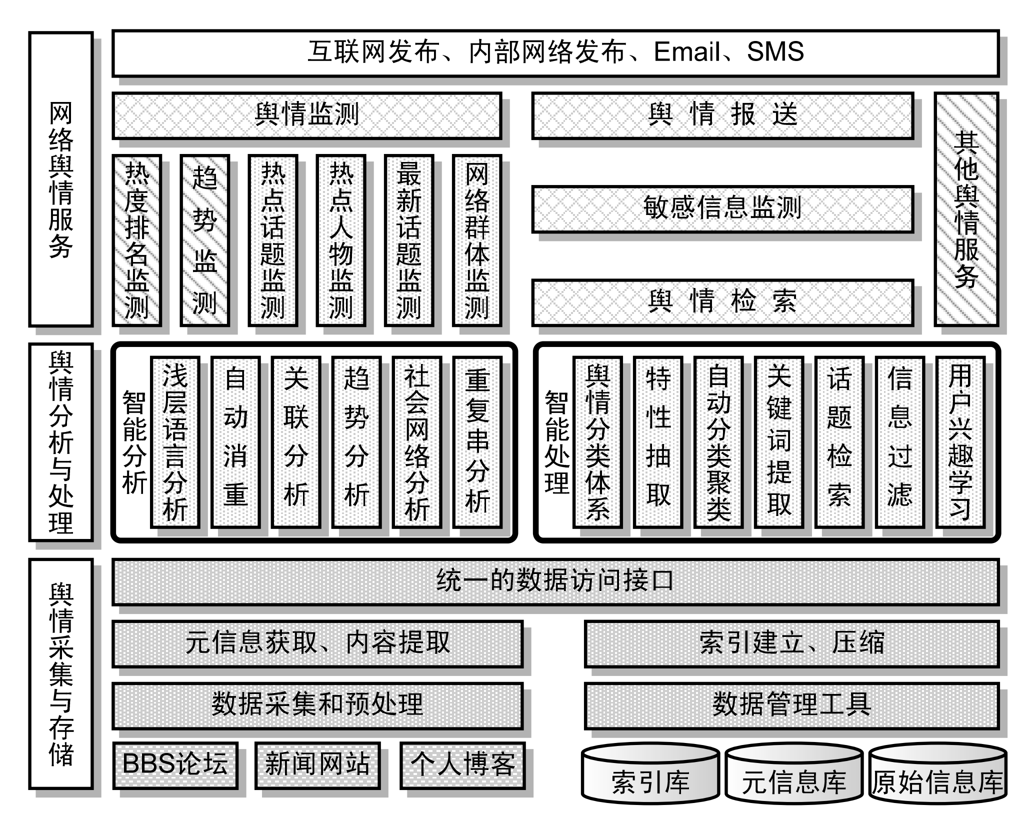


图1. 天玑舆情系统基本架构

天玑舆情监测系统是专门针对网络舆情监测的工作要求和特点而打造的，不仅采用了专业化的搜索引擎技术，还融入了更加智能的数据挖掘技术，建立一个以日为周期的网络舆情

监测平台，同时配上以周或者以月为基础的舆情分析报告，从而提供了一个便捷、科学、可操作的舆情工作平台。

天玑舆情系统采用的信息采集、信息抽取、汉语分词、全文检索、自动消重、关联分析、关键词提取、自动摘要、分类聚类、个性化信息主动推送等技术，能够对海量信息进行智能检索、智能处理、智能分析，为用户提供热点专题监控、个性化信息推送、精确全文检索等服务，提供客观、完整、准确的舆情监测报告，便于用数据、用图表来解读舆情，减少主观臆断。

天玑舆情系统采用图 1 所示的技术和三层软件架构，能满足舆情监测的以下需求：

1. 提高效率，扩大了解民情舆情的范围

人工监控网络信息不仅速度慢、效率低，要花费大量的人力和物力，并且还不能穷尽所关注的热点和专题。而利用天玑系统的信息采集技术，可以 24 小时不间断地监控全网和成千上万个指定频道的信息，并且不会重复和遗漏。无疑增加了舆情监控的数量和范围，提高了品质。

2. 舆情监测实时准确

由于采取了 24 小时定向监控，网络任何的最新信息都能被及时发现和识别，通过监控知识库的判断和分析，一发现问题就产生报警消息，及时推送给管理部门知晓，使之及时掌握舆情爆发点和事态发展趋势，真正起到了“网络监督员”的作用。

3. 突发事件从容应对

遇到突发事件和重大专题，系统自动进行首发地址的追踪、统计走势和传播路径的分析，并能够 24 小时不间断地对删帖状态和网络活跃分子进行监控，对网民的多种观点和意见进行分析，从而能够有效把握舆情态势，并通过权限联动产生各方面力量的合力。

4. 灵活生成证据库和舆情报告

系统能够对关键信息生成证据库进行永久保存；并能够做到日日简报，月月专报，要事快报；还可以针对社会舆论热点，自动生成热点舆情报告。报告内容图文并茂，提供了大量的统计分析和网民观点分析，能够为辅助决策提供很大帮助。任务完成后可生成处置结果报告，反映工作的效果和效率。

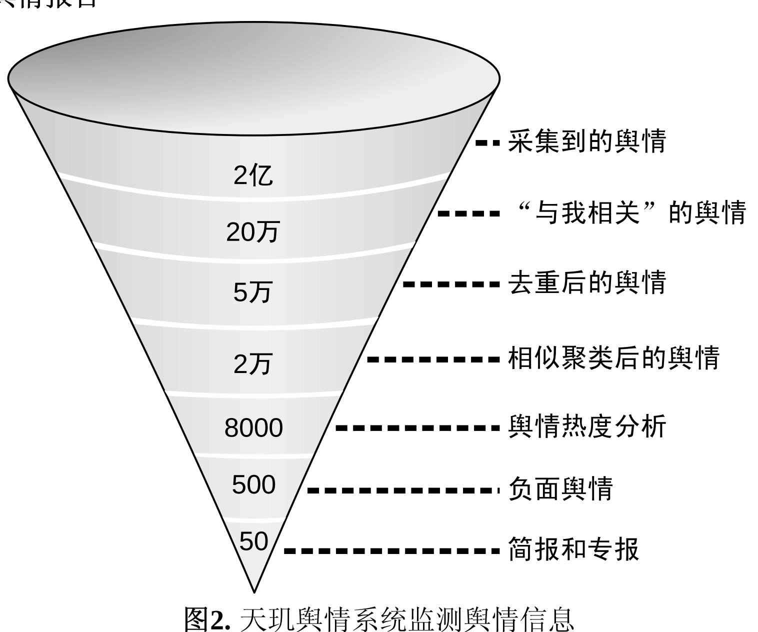


图2. 天玑舆情系统监测舆情信息

在天玑舆情系统的协助下，仅仅关注系统智能推送的少量信息，就能达到全网监测的效果，舆情工作从此变得高效而有趣。

5 天玑舆情系统特点

天玑网络舆情监测系统融合了中科院计算所在内容深度挖掘技术领域 10 多年来的研究成果，含有多项专利技术，其前期的互联网智能搜索挖掘系统曾获得国家科技进步一等奖。其中的多文档摘要、网页与博客专家搜索、信息过滤 3 项技术先后获得了信息智能处理世界级比赛——国际文本检索（TREC）大赛第一名；中文分词技术国内外公开测评第一，是国际上公认的汉语分析第一品牌；话题发现与跟踪技术获得国际话题检测与跟踪（Topic Detect and Tracking, TDT）评测全球第二名等多项荣誉。先进的技术带来了天玑舆情系统全面、高效、准确、深入等四大优点，具体体现在以下四个方面：

1. 信息全面获取

信息获取是指从网络中高速准确地采集数据，提取相关元信息。天玑网络舆情监测系统能够快速高效获取网页，支持 Javascript 等多种复杂页面形式，支持网页编码自动识别和转换，支持基于 cookie 状态检测的采集，具有反“防刷新”的采集机制；可定向采集并抽取新闻、论坛、博客等各类复杂的信息内容；支持元搜索主题采集，在各大搜索引擎基础上只采集用户感兴趣的内容，信息全面，更新及时。

2. 自然语言理解

自然语言理解的目标是让计算机像人类一样真正理解各类庞杂信息中的语言语义，为进一步的深入挖掘提供可信的知识依据。天玑舆情系统内嵌强大的汉语词法分析器 ICTCLAS，集成了高效的正文与关键词提取技术，可去除网页中的噪音，自动计算出有代表性的关键词汇；文本分类聚类算法快速精准；文档摘要可以自动分析文档的内容，提供简短准确的文本摘要。

3. 信息智能搜索

信息智能搜索可提供更智能化、专业化与人性化的信息搜索服务。天玑舆情系统采用分布式全文检索系统 I3Search，嵌入了查询理解的最新研究成果，自动挖掘潜在语义关联，内核经过精心设计，是高扩展性与高性能的完美组合。系统支持文本、数字、日期、字符串等数据类型的高效索引；支持丰富的查询语言，同时支持 32 位与 64 位硬件平台下的 Windows、Linux 等主流操作系统。索引速度高达 9M/s，支持在线索引，实现毫秒级别查询。

4. 舆情综合挖掘

提供被监测信息源的一站式、全方位的监控和浏览。

舆情综合挖掘面向互联网日益增长的舆情监测、竞争情报与危机公关需求，对从互联网上采集到的论坛、博客、新闻、搜索引擎、新闻评论、跟贴、图片、音视频等信息，综合挖掘分析，实现话题的自动发现和全方位跟踪、溯源，提供时间、空间分布及趋势分析；对文章、评论的倾向性进行智能分析；深入挖掘网络对象之间的关系；推送有价值的舆情信息和统计报表，提供舆情监测与危机公关应对服务。

天玑网络舆情监测系统架构具有良好的可扩展性，可根据需求灵活定制，已在政府、金融、教育等行业用户取得了良好的应用效果。天玑系统已经广泛地应用于工信部、广电总局、证监会等关键部门，发挥了实际作用。其中，在中国证监会建设的网络信息监控系统使得证监会的舆情监控工作取得了长足的进步，为维护资本市场的稳定，保护中小投资者的利益，提供了有效的支持，获得了 2009 年度证券期货业科学技术奖二等奖。

参考文献

[1] 中国互联网信息中心 2009 年 7 月*第 24 次中国互联网发展状况统计报告*[OL].

[2] 王来华.舆情研究概念——*理论、方法和现实热点*[M].天津:天津社会科学院出版社，2003.

[3] 中共中央宣传部舆情信息局.*舆情信息工作概论*[M].北京:学习出版社，2006.

[4] 楼玲娣，周小斌.*网络舆情的运行状态分析*[J].特区实践与理论，2009.2.

作者简介：

翟立东： 中国科学院计算技术研究所 网络重点实验室 助理研究员

余智华： 中国科学院计算技术研究所 网络重点实验室副主任 副研级高级工程师
Email: yzh@ict.ac.cn

丁国栋： 中国科学院计算技术研究所 网络重点实验室 副研级高级工程师

（ 上接第 15 页 ）

[21] Zhou, G. and J. Su, *Named entity recognition using an HMM-based chunk tagger, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* 2002, Association for Computational Linguistics: Philadelphia, Pennsylvania. p. 473-480.

[22] Liu, D.C. and J. Nocedal, *On the limited memory BFGS method for large scale optimization*. Math. Program., 1989. 45(3): p. 503-528.

[23] Vishwanathan, S.V.N., et al., *Accelerated training of conditional random fields with stochastic gradient methods, in Proceedings of the 23rd international conference on Machine learning.* 2006, ACM: Pittsburgh, Pennsylvania. p. 969-976.

[24] Hobbs, J.R., *The generic information extraction system, in Proceedings of the 5th conference on Message understanding.* 1993, Association for Computational Linguistics: Baltimore, Maryland. p. 87-91.

[25] H. Cunningham, D.M., K. Bontcheva, V. Tablan. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. . in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. July 2002. Philadelphia..

[26] Gaizauskas, R., et al., *University of Sheffield: description of the LaSIE system as used for MUC-6, in Proceedings of the 6th conference on Message understanding.* 1995, Association for Computational Linguistics: Columbia, Maryland. p. 207-220.

[27] T. S. Jayram, R.K., Sriram Raghavan, Shivakumar Vaithyanathan, Huaiyu Zhu, *Avatar information extraction system*. IEEE Data Engineering Bulletin, 2006. 29: p. 40–48.

作者简介：

孟　涛　中国科学院计算技术研究所、助理研究员，Email: mengtao@software.ict.ac.cn

曹　雷　中国科学院计算技术研究所、博士研究生

折闪电　中国科学院计算技术研究所、硕士研究生

程学旗　中国科学院计算技术研究所、研究员